

Chromosomal assembly of the Antarctic toothfish (*Dissostichus mawsoni*) genome using third-generation DNA sequencing and Hi-C technology

DEAR EDITOR,

The Antarctic toothfish, *Dissostichus mawsoni*, belongs to the Nototheniidae family and is distributed in sub-zero temperatures below 60° latitude in the Southern Ocean. Therefore, it is an attractive model species to study the stenothermal cold-adapted character state. In this study, we successfully generated highly contiguous genome sequences of *D. mawsoni*, which contained 1 062 scaffolds with a N50 length of 36.98 Mb and longest scaffold length of 46.82 Mb. Repetitive elements accounted for 40.87% of the genome. We also inferred 32 914 protein-coding genes using *in silico* gene prediction and transcriptome sequencing and detected splicing variants using Isoform-Sequencing (Iso-Seq), which will be invaluable resource for further exploration of the adaptation mechanisms of Antarctic toothfish. This new high-quality reference genome of *D. mawsoni* provides a fundamental resource for a deeper understanding of cold adaptation and conservation of species.

Many unique fish live in the Southern Ocean surrounding Antarctica within the coldest waters on Earth. It has been isolated by the Antarctic circumpolar current (Eastman, 2005; Livermore et al., 2005) where sea temperatures range around the ice point (−1.9 °C) for most of the year. Antarctic fish, which include cold-adapted teleosts, are dominated by a single lineage belonging to the Perciformes suborder Notothenioidae. This suborder consists of eight families and >100 species and accounts for ~90% of total fish biomass in the Antarctic Ocean (Eastman & De Vries, 1981). From a common ancestor, a variety of closely related species with distinct ecological status as well as size, shape, and color have emerged in the Southern Ocean. Therefore, genomic research is essential to understand the environmental adaptation and evolution of these fish.

Open Access

This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright ©2021 Editorial Office of Zoological Research, Kunming Institute of Zoology, Chinese Academy of Sciences

The Antarctic toothfish, *Dissostichus mawsoni*, belongs to the family Nototheniidae of the order Perciformes and is native to the Southern Ocean. It is distributed below 60° latitude and is the largest of all Antarctic fish (2.0 m in length and 140 kg in mass) (Eastman & De Vries, 1982). Its stenothermal cold-adapted state makes the species an attractive model for evolutionary and genomic studies among Antarctic fish. The Antarctic toothfish is an economically important fishery species, with a commercial catch in Subarea 88.1 of 2 680 tons in 2018 (Maschette et al., 2019).

Recently, *de novo* assembly of the Antarctic toothfish genome and extensive transcriptomic characterization using short-read Illumina data have been reported (Chen et al., 2019), though the genome was fragmented into many scaffolds due to sequencing by synthesis technology. The development of third-generation single-molecule sequencing technology has enabled the production of long-read sequences and the discovery of the features of previously unavailable DNA regions. Here, we report on a re-assembled whole-genome of *D. mawsoni* using long-read sequencing and Hi-C technology, which should help provide comprehensive insight into its adaptive mechanisms.

Antarctic toothfish (length ~50 cm, sex not determined) were collected using a vertical setline in the eastern Ross Sea (Subarea 88.1), Antarctica (<http://www.fao.org/fishery/area/Area88/en>), during the austral summer season (December 2018). Specimens were killed for tissue sampling and then rapidly frozen for further analysis. All sample collection and experimental protocols were in compliance with the laws regarding activities and environmental protection in Antarctic and were approved by the Minister of Foreign Affairs and Trade of the Republic of Korea.

To obtain sufficient high-quality DNA molecules for the

Received: 14 September 2020; Accepted: 30 November 2020; Online: 01 December 2020

Foundation items: This study was supported by a grant from the National Institute of Fisheries Science (NIFS) of the Republic of Korea (R2019021) and “Ecosystem Structure and Function of Marine Protected Area (MPA) in Antarctica” project (PM19060) funded by the Ministry of Oceans and Fisheries (20170336), Korea

DOI: 10.24272/j.issn.2095-8137.2020.264

PacBio Sequel platform (Pacific Biosciences, USA), one *D. mawsoni* fish was dissected and muscle tissue was used for DNA extraction using the phenol/chloroform extraction method. DNA quality was checked using a fragment analyzer system (Agilent Technologies, USA) and Qubit 2.0 fluorometer (Invitrogen, Life Technologies, USA). The single-molecule real-time (SMRT)bell library was sequenced using eight SMRT cells (Pacific Biosciences, Sequel™ SMRT Cell 1M v2) with a Sequel Sequencing Kit 2.1 (Pacific Biosciences, USA) and 1×600 min movies were captured for each SMRT cell using the Sequel sequencing platform (Menlo Park, USA). The average coverage of the SMRT sequences was ~81-fold (Supplementary Table S1).

Muscle tissue from the same sample was used to construct a Hi-C chromatin contact map for chromosome-level assembly. Tissue fixation, chromatin isolation, and library construction were performed according to the manufacturer's instructions (Dovetail Genomics, USA) (Belton et al., 2012). After checking the insert size, concentration, and effective concentration of the constructed libraries, the final libraries were sequenced using the Illumina NovaSeq 6000 platform (San Diego, USA) with a 150-bp paired-end strategy. A total of 874 million raw reads were generated from the Hi-C libraries (Supplementary Table S1) and were mapped to the polished *D. mawsoni* contigs using HiC-Pro (v2.8.0) with default parameters.

For transcriptome sequencing, we prepared 1 µg of pooled total RNA from the muscle and skin of *D. mawsoni*. Using a SMARTer PCR cDNA Synthesis Kit (Clontech, USA), RNA was synthesized to cDNA. The SMRTbell library was constructed using the SMRTbell™ Template Prep Kit 1.0-SPv3. The SMRTbell library was sequenced using SMRT cells (Pacific Biosciences, Sequel™ SMRT Cell 1M v2) and the Sequel Sequencing Kit 2.1. For each SMRT cell, 1×600 min movies were captured with the use of the Sequel sequencing platform; the pre-extension time was 240 min (Supplementary Table S1). The Iso-Seq sequencing data were analyzed using SMRT Link (v6.0.0).

For *de novo* genome assembly, the FALCON-Unzip assembler (v0.4, Falcon, RRID:SCR 016089) was used (Chin et al., 2013) with parameters of `length_cutoff=12 000` and `length_cutoff_pr=10 000` and with filtered subreads from SMRT Link (v4.0.0) (minimum subread length=50). To improve the quality of the genome assembly, the FALCON-Unzip assembler was polished using the Arrow algorithm with unaligned BAM files as raw data.

A draft *D. mawsoni* genome was previously generated using Illumina short-read sequencing (Chen et al., 2019). However, since several gaps prevailed in the draft genome set and there was no information about the linkage group, it was difficult to compare the structure of the Antarctic toothfish genome at the chromosomal scale. To improve this genome resource, long-read SMRT sequencing from Pacific Biosciences and Hi-C scaffolding were implemented. First, we performed *de novo* assembly of the long PacBio sequence reads using the FALCON-Unzip tool and obtained a genome assembly with a

size of 924.75 Mb, an N50 contig size of 3.23 Mb, and longest contig size of 24.49 Mb (Supplementary Table S2). To construct the reference genome at the chromosome level, we constructed a Hi-C library and anchored the scaffolds into chromosomes after quality control using the HiC-Pro, Juicer (v1.5) (Durand et al., 2016) and 3D-DNA (v170123) pipeline (Dudchenko et al., 2017) based on the draft genome assembly (Figure 1A). The assembled genome was 926.3 Mb (GC content: 41.57%) in length with a scaffold N50 of 36.98 Mb and longest scaffold of 46.82 Mb. In total, there were 1 062 scaffolds in the *D. mawsoni* genome assembly, with 24 chromosome-scale scaffolds occupying 91.3% of the assembly (Supplementary Tables S2, S3 and Figure S1).

Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0 (Simão et al., 2015) (RRID:SCR 015008) was used along with the actinopterygii odb9 database to assess the completeness of the new *D. mawsoni* genome assembly. Among 4 584 BUSCO groups searched, 4 197 and 194 BUSCO core genes were completely and partially identified, respectively, contributing to a total of 95.7% BUSCO genes in the *D. mawsoni* genome (Supplementary Table S4).

The diploid chromosome number (2n) of *D. mawsoni* is 48 (Ghiogliotti et al., 2007). Comparison of its chromosome-scale assemblies with those of the *Gasterosteus aculeatus* genome (2n=42) showed a highly similar synteny (Figure 1B). However, each of three chromosomes (from Groups 1, 4, and 7) of *G. aculeatus* matched with two chromosomes in *D. mawsoni* (scaffolds 23 and 35, scaffolds 13 and 22, scaffolds 11 and 24, respectively) (Supplementary Figure S2).

A *de novo* repeat library was constructed using RepeatModeler (v1.0.3) (Bao & Eddy, 2002), which included RECON (RRID:SCR 006345) and RepeatScout v1.0.5 (RRID:SCR 014653) (Price et al., 2005) software with default parameters. The Tandem Repeats Finder (Benson, 1999) program was used to predict the consensus sequences and classification information for each repeat. We analyzed the repetitive sequences in the *D. mawsoni* genome, including those in the tandem repeats and transposable elements (TEs). TEs play an important role in shaping genome architecture and are a source of regulatory mutations in evolution. A difficulty arises in representing TEs in genome assemblies based on short Illumina sequence reads. Therefore, our long-read sequences greatly improved both the length and quantity of the TE repeats in the *D. mawsoni* genome assembly compared to the published short-read assembly. The *D. mawsoni* genome contained 40.87% of repeat sequences, including 36.27% (336 Mb) of TEs such as long terminal repeats (LTRs, 4.21%), short interspersed nuclear elements (SINES, 0.50%), long interspersed nuclear elements (LINEs, 5.35%), and DNA transposons (15.51%) (Supplementary Table S5 and Figure S3). Divergence of TEs was examined using RepeatMasker software, where Kimura distances (Kimura, 1980) estimated for aligned TEs (`calcDivergenceFromAlign.pl`) were used to draw repeat landscapes (`createRepeatLandscape.pl`). The *D. mawsoni* genome had a higher number of recent TE insertions (Kimura

divergence K -values ≤ 5) that were strongly shaped by DNA transposons (Supplementary Figure S4). Because K -values calculated for TEs can reflect age and transposition history (Chalopin et al., 2015), we concluded that there have been recent transposable element bursts in the Antarctic toothfish.

Genome annotation was conducted using MAKER v2.28 (RRID:SCR_005309) (Holt & Yandell, 2011), which is a portable and easily configurable genome annotation pipeline. Subsequently, repeat masked genomes were used for *ab initio* gene prediction with SNAP v2006-07-28 (SNAP, RRID:SCR_002127) (Korf, 2004) and Augustus (Augustus: Gene Prediction, RRID:SCR_008417) software. MAKER was initially run in the est2genome mode based on the Iso-Seq data, including 57 406 full-length transcripts. Additionally, protein evidence was obtained from the genomes of 19 teleosts, including three Antarctic fish (Supplementary Table S6). Exonerate software, which provides integrated information for the SNAP program, was used to polish MAKER alignments. MAKER was then used to select and revise the final gene model considering all available information. Other non-coding RNAs in the Antarctic toothfish assembly were identified using Infernal (v1.1) (Nawrocki & Eddy, 2013) and covariance models (CMs) from the Rfam database v12.1 (Rfam, RRID:SCR_007891) (Griffiths-Jones et al., 2005). Putative tRNA genes were identified using tRNAscan-SE v1.3.1 (tRNAscan-SE, RRID:SCR_010835) (Lowe & Eddy, 1997), which uses a CM that scores candidates based on their sequences and predicted secondary structures.

The predicted genes were aligned to the NCBI non-redundant protein (nr) (Benson et al., 1999), SwissProt (RRID:SCR_002380) (Boeckmann et al., 2003), TrEMBL (RRID:SCR_002380) (Boeckmann et al., 2003), KOG (EuKaryotic Orthologous Groups) (Tatusov et al., 2001), and KEGG (Kyoto Encyclopedia of Genes and Genomes, RRID:SCR_001120) (Kanehisa & Goto, 2000) databases using BLAST v2.2.31 (Altschul et al., 1990) with a maximum e -value of $1e-5$. Gene Ontology (GO) (RRID:SCR_002811) terms (Dimmer et al., 2012) were assigned to the genes using the Blast2Go v4.0 pipeline (RRID:SCR_005828) (Conesa et al., 2005).

A total of 32 914 protein-coding genes in the *D. mawsoni* genome were annotated using a combination of *ab initio* gene prediction, homology searching, and transcript mapping. The coding sequence comprised 51.2 Mb (exons 55.2 Mb) with an average of 9.7 exons per gene (Supplementary Table S7). Consequently, a total of 20 202 genes were annotated in >1 database (Supplementary Table S7). A total of 24 920, 19 205, and 14 474 genes were annotated in the GO, KOG, and KEGG databases, respectively, and the functional classifications of these genes are presented in Supplementary Figures S5–S7.

We identified orthologous gene clusters using the OrthoMCL (Li et al., 2003) pipeline, which applied the Markov Clustering Algorithm (MCL) with default options in all steps for the genome sequences of the 20 species (Supplementary Table S6). It was critical for analysis to include representative

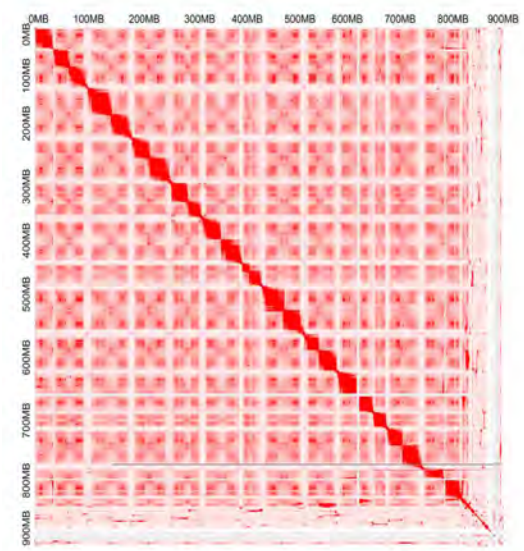
species of diverse phylogenetic clades and the 20 species were selected among those with well-annotated and well-assembled genomes.

Phylogenetic tree construction was performed based on single-copy orthologous genes. The sequences of protein-coding genes were aligned using a Probabilistic Alignment Kit (PRANK) (Löytynoja & Goldman, 2005) with the codon alignment option. The maximum-likelihood method was applied to construct a phylogenetic tree using RAXML with 1 000 bootstraps, and divergence times were calibrated with TimeTree (median estimates of pairwise divergence time for *D. rerio* and *G. morhua*: 230.4 million years ago) (Hedges et al., 2006).

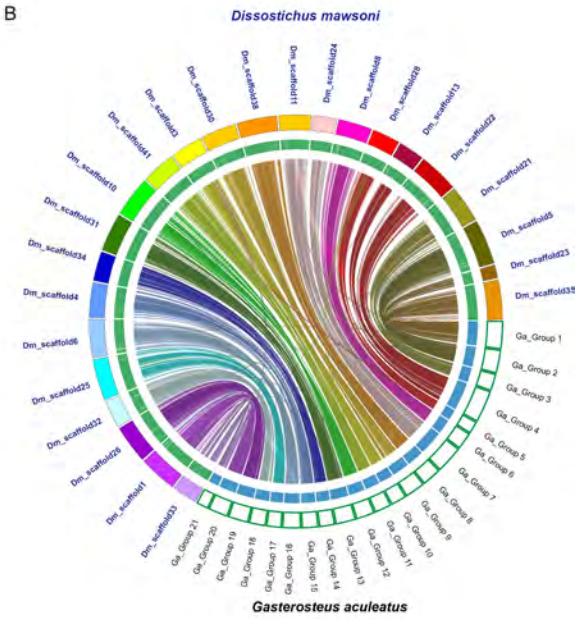
Ortholog gene families from each species were identified: 7 731 orthologous gene families were commonly identified among the 20 teleosts, including 434 (1 431 genes) paralogous gene families that were *D. mawsoni*-specific (Supplementary Table S8). The phylogenetic tree of *D. mawsoni* and the 19 teleost species was constructed using 1 422 single-copy orthologs (Figure 1C). Among the 20 fish species, *D. mawsoni* and three other Antarctic fish were clustered together on the branch of a non-Antarctic fish, *G. aculeatus*, with a divergence time of around 105 million years ago. Furthermore, *D. mawsoni* diverged approximately 28 million years ago from the Antarctic fish *Chaenocephalus aceratus*. Analysis of gene gain-and-loss among genomes enables the reconciliation of a species tree with the gene tree for each family. Here, *D. mawsoni* had 659 (including 2 114 genes) significantly expanded and 116 (including 136 genes) significantly contracted gene families (Figures 1C, D). The vast majority of the expanded biological pathways belonged to two functional categories: (i) involved in nervous system functions (neuron projection development, GO:0031175; neuron development, GO:0048666; cell morphogenesis involved in neuron differentiation, GO:0048667; generation of neurons, GO:0048699; neuron projection morphogenesis, GO:0048812; axon development, GO:0061564) and (ii) cellular component morphogenesis (cell projection organization, GO:0030030; cell part morphogenesis, GO:0032990; cell projection morphogenesis, GO:0048858). In the molecular function category, peptidase regulator activity (endopeptidase regulator activity, GO:0061135; peptidase inhibitor activity, GO:0030414; endopeptidase inhibitor activity, GO:0004866), and signaling receptor binding (endopeptidase inhibitor activity, GO:0004866) were the major expanded pathways (Supplementary Tables S9, S10). In addition, 14 055 orthologous gene families containing 16 162 genes in *D. mawsoni* were commonly identified in the four Antarctic fish. Moreover, 621 gene families were *D. mawsoni* species-specific paralogs (Figure 1E) involved in DNA metabolic processes (DNA biosynthetic process, GO:0071897; DNA integration, GO:0015074; RNA-dependent DNA biosynthetic process, GO:0006278) (Supplementary Table S11).

Splicing variants were analyzed using SQANTI2 (Tardaguila et al., 2018) with the Iso-Seq data as full-length transcript

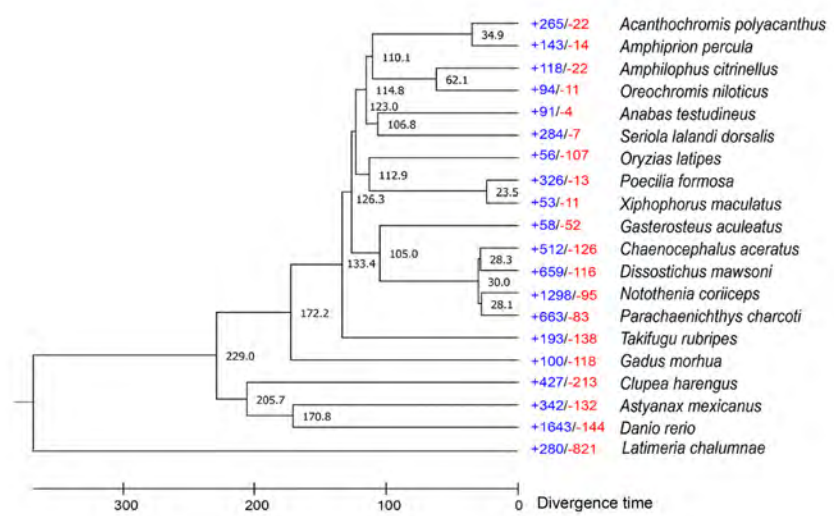
A



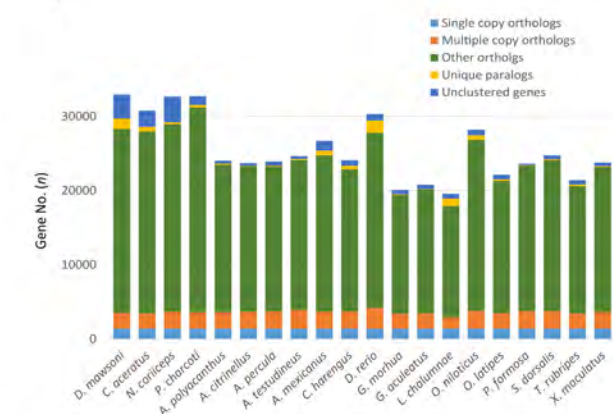
B



C



D



E

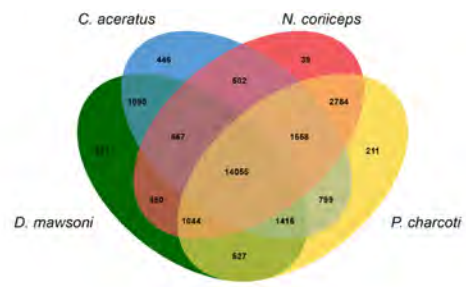


Figure 1 Genome analysis of *D. mawsoni*

A: Hi-C interaction heat map for *D. mawsoni* genome. Intensity of red dots indicates density of cross-linking information between genome fragments. Higher intensity was generated in the same chromosome. B: Collinear relationship between *D. mawsoni* and *G. aculeatus*. Green boxes represent chromosomes of *G. aculeatus* and multicolored ones represent chromosomes of *D. mawsoni*. C: Phylogenetic analysis of *D. mawsoni* within teleost lineage and gene family gain-and-loss analysis including number of gained (+) and lost gene families (-). Black numbers specify divergence times between lineages. D: Orthologous gene families among *D. mawsoni* and other fish species. E: Venn diagram of orthologous gene families among *D. mawsoni* and three Antarctic fish species.

sequences. The Iso-Seq data were aligned to the assembled genome using Minimap2 (Li, 2018) and the collapsed high-quality isoforms were aligned to unique isoforms using the Cupcake ToFU pipeline (Tseng, 2017). SQANTI2 extracted various types of splicing variants using the collapsed isoforms and the gene prediction information of the assembled genome. After excluding novel, mono exonic, and antisense transcripts, genes were selected based on the ascending order of the number of isoforms. Enrichment analysis of genes of splicing variants was performed using the Blast2GO v4.0 pipeline (RRID:SCR_005828) (Conesa et al., 2005) with FDR<0.5.

Iso-Seq data analysis identified 31 480 unique isoforms in 14 565 unique genes. Most novel genes were located in the intergenic region (Supplementary Tables S12, S13). Functional annotation using enrichment analysis by Fisher's Exact Test for genes with more than 10 splicing variants (Supplementary Table S14) identified genes related to development, anatomical structure development, functional annotations (system development, GO:0048731; animal organ development, GO:0048513; tissue development, GO:0009888; cell development, GO:0048468; embryo development, GO:0009790; muscle structure development, GO:0061061; epithelium development, GO:0060429; and circulatory system development, GO:0072359), and organization related function (cytoskeleton organization, GO:0007010; protein-containing complex subunit organization, GO:0043933; actin cytoskeleton organization, GO:0030036; supramolecular fiber organization, GO:0097435; and organelle organization, GO:0006996) (Supplementary Table S15).

In the current study, we presented a high-quality chromosome-level genome assembly of the Antarctic toothfish, *D. mawsoni*, using PacBio Sequel sequencing and Hi-C chromatin contact maps. The *D. mawsoni* genome assembly (926 Mb) included 24 chromosomes that accounted for 91% (840 Mb) of all genome sequences. The *D. mawsoni* genome contained 32 914 protein-coding genes and 434 paralogous *D. mawsoni*-specific gene families among 20 teleost fish and 621 paralogous *D. mawsoni*-specific gene families among the four Antarctic teleost fish. This chromosome-length genome assembly will not only provide insights into the molecular and ecological adaptations of Antarctic fish to extreme environments but will also facilitate exploration of genomic adaptations to a wide range of evolutionary, ecological, metabolic, developmental, and biochemical features of Antarctic fish.

DATA AVAILABILITY

The *Dissostichus mawsoni* genome project was deposited in NCBI under BioProject No. PRJNA574770 and the Whole-Genome Shotgun project was deposited at DDBJ/ENA/GenBank under accession No. JAAKFY000000000. The version described in this paper is JAAKFY010000000. The genome browser, assembly, and annotation data are accessible on <http://genome.kusglab.org/>.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

H.P. and H.-W.K. conceived the study. S.J.L., E.J., S.-G.C., S.C. E.C., J.K., and H.P. performed genome sequencing, assembly, and annotation. E.J. and S.J.L. performed experiments. S.J.L., J.-H.K., H.-W.K., and H.P. mainly wrote the manuscript. All authors contributed to writing and editing the manuscript as well as collating the supplementary information and creating the figures. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We would like to thank Sunwoo Corporation for providing the Antarctic toothfish samples.

Seung Jae Lee^{1,†}, Jeong-Hoon Kim^{2,†}, Euna Jo^{1,3},
Eunkyung Choi¹, Jinmu Kim¹, Seok-Gwan Choi⁴,
Sangdeok Chung⁴, Hyun-Woo Kim^{5,*}, Hyun Park^{1,*}

¹ College of Life Sciences and Biotechnology, Korea University,
Seoul 02841, Republic of Korea

² Division of Polar Life Science, Korea Polar Research Institute,
Incheon 21990, Republic of Korea

³ Unit of Research for Practical Application, Korea Polar Research
Institute, Incheon 21990, Republic of Korea

⁴ National Institute of Fisheries Science (NIFS), Busan 46083,
Republic of Korea

⁵ Department of Marine Biology, Pukyong National University,
Busan 48513, Republic of Korea

[†]Authors contributed equally to this work

*Corresponding authors, E-mail: kimhw@pknu.ac.kr;
hpark@korea.ac.kr

REFERENCES

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3): 403–410.
- Bao ZR, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, **12**(8): 1269–1276.
- Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**(3): 268–276.
- Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BFF, Rapp BA, et al. 1999. GenBank. *Nucleic Acids Research*, **27**(1): 12–17.
- Benson GJ. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**(2): 573–580.
- Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**(1): 365–370.
- Chalopin D, Naville M, Plard F, Galiana D, Volff JN. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, **7**(2): 567–580.
- Chen LB, Lu Y, Li WH, Ren YD, Yu MC, Jiang SW, et al. 2019. The genomic basis for colonizing the freezing Southern Ocean revealed by Antarctic toothfish and Patagonian robalo genomes. *GigaScience*, **8**(4): giz016.
- Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, **10**(6): 563–569.
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**(18): 3674–3676.
- Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, et al. 2012. The UniProt-GO annotation database in 2011. *Nucleic Acids Research*, **40**(D1): D565–D570.
- Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**(6333): 92–95.
- Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, **3**(1): 95–98.
- Eastman JT. 2005. The nature of the diversity of Antarctic fishes. *Polar Biology*, **28**(2): 93–107.
- Eastman JT, DeVries AL. 1981. Buoyancy adaptations in a swim-bladderless Antarctic fish. *Journal of Morphology*, **167**(1): 91–102.
- Eastman JT, DeVries AL. 1982. Buoyancy studies of notothenioid fishes in McMurdo Sound, Antarctica. *Copeia*, **1982**(2): 385–393.
- Ghigliotti L, Mazzei F, Ozouf-Costaz C, Bonillo C, Williams R, Cheng CHC, et al. 2007. The two giant sister species of the Southern Ocean, *Dissostichus eleginoides* and *Dissostichus mawsoni*, differ in karyotype and chromosomal pattern of ribosomal RNA genes. *Polar Biology*, **30**(5): 625–634.
- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, **33**(S1): D121–D124.
- Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**(23): 2971–2972.
- Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**(1): 491.
- Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**(1): 27–30.
- Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**(2): 111–120.
- Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics*, **5**(1): 59.
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18): 3094–3100.
- Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**(9): 2178–2189.
- Livermore R, Nankivell A, Eagles G, Morris P. 2005. Paleogene opening of Drake passage. *Earth and Planetary Science Letters*, **236**(1–2): 459–470.
- Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**(5): 955–964.
- Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(30): 10557–10562.
- Maschette D, Wotherspoon S, Ziegler P. 2019. Exploration of CPUE Standardisation Variances in the Ross Sea (Subareas 88.1 and 88.2A South of 70°S) Antarctic Toothfish (*Dissostichus mawsoni*) Exploratory Longline Fishery. Hobart, Tasmania: CCAMLR.
- Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**(22): 2933–2935.
- Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics*, **21**(S1): i351–i358.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**(19): 3210–3212.
- Tardaguila M, De La Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research*, **28**(3): 396–411.
- Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, et al. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, **29**(1): 22–28.
- Tseng E. 2017(2020-10-16). Cupcake ToFU: supporting scripts for Iso-Seq after clustering step. https://github.com/Magdoll/cDNA_Cupcake/wiki/Cupcake-ToFU-supporting-scripts-for-Iso-Seq-after-clustering-step.